Capitation and Financial Risk Allocation in New Zealand’s Primary Health Care Sector: The Perverse Consequences of Neglecting Financial Risk Allocation

Bronwyn Howell1

Abstract
Using analysis of the management of ‘random’ and ‘controllable’ risk in capitation contracts, this paper critiques the arrangements in the New Zealand Primary Health Care Strategy (NZPHCS) introduced in 2002. Total system costs have undoubtedly risen under the mixed capitation model adopted, relative to fee-for-service. By requiring only those treated to pay all costs not factored into the government’s prospective capitation payments, the burden of unanticipated risk-management costs falls disproportionately on the sickest patients. Rather than resources being allocated on the basis of health need, the sickest individuals bear a disproportionate share of the cost of random demand shocks.

Introduction
There are many mechanisms for paying physicians; some are good and some are bad. The three worst are fee-for-service, capitation and salary. (Robinson, 2001:149)

Partial or full capitation contracts have become common in primary health care remuneration; for example, in England’s Primary Care Trusts (Keen, Light and Mays, 1999), United States Managed Care schemes (Robinson, 2004; Hagen, 1999) and New Zealand’s Primary Health Care Strategy (King, 2001). Capitation is also used extensively in public sector ‘outcomes-based’ (Honore et al., 2004) and ‘performance-based’ (Martin, 2002) contracting. Reduced emphasis on the consultation as the primary payment determinant is attributed with shifting primary care delivery focus away from interventions in the event of illness towards the promotion and maintenance of wellness (Coster and Gribben, 1999; Cumming, 1999; Malcolm, 1997). Capitation is also attributed with stimulating

1 New Zealand Institute for the Study of Competition and Regulation Inc. and Victoria Management School, Victoria University of Wellington, PO Box 600, Wellington, New Zealand. Email bronwyn.howell@vuw.ac.nz The author wishes to acknowledge the helpful comments provided by Lewis Evans, the support of Glenn Boyle and the New Zealand Institute for the Study of Competition and Regulation, whilst undertaking research into New Zealand primary health care markets, and the Deane Endowment Trust for financial support during the preparation of this paper. Any errors or omissions remain the responsibility of the author.
increased equity, targeting high health need, encouraging a team approach to primary health care, and a change in focus towards a care management model as opposed to a model of episodic intervention with a focus on illness (Crampton, Sutton and Foley, 2001).

However, capitation contracts also have significant limitations. Robinson (2001:4) identifies that capitation is inferior to fee-for-service in that it does not recognise the extent of practitioner effort exerted: “Its payment is determined prospectively without regard to the number of services provided, overpaying physicians who stint on care and underpaying those who provide many complex services.” Capitation also performs poorly in regard to risk management, he says, as it is “imperfectly adjusted for the severity of illness of each covered patient. Even a well-adjusted capitation payment rate fails to compensate physicians who treat patients whose condition deteriorates, leading to greater utilization and cost, for reasons independent of the physician’s own actions.”

This paper explores the ways in which the financial consequences of unpredictable events affecting demand for services are allocated under the capitated contracts used in the New Zealand Primary Health Care Strategy (NZPHCS). Section one discusses the general effect that unpredictable events can have upon capitated practitioners’ incomes. Section two then discusses the particular arrangements of the New Zealand capitation contracts introduced in 2002. The perverse effects arising from the New Zealand contracts are illustrated in section three using two recent exogenous events: a strike by junior doctors in the country’s hospitals, and a decision by District Health Boards (DHBs) to remove all individuals on waiting lists for secondary and tertiary (hospital-provided) consultations and procedures for six months or longer back to their primary care providers for ongoing management. Section four concludes.

Capitation in Primary Health Care Markets

In a partial capitation contract, the remunerated party’s income comprises a fixed component $f$ determined by the ex-ante characteristics of the population for whom care responsibility is assumed, and a variable component $v$ for each unit of output $q$ (for example, consultations) produced at a cost $c$ (profit = $f - q(c-v)$). In a full capitation contract ($v = 0$, equivalent to budget funding of entities or salaried remuneration of individuals), the recipient’s income is invariant to the number of consultations provided. By decoupling remuneration from cost drivers, the higher is $f$ and the lower is $v$, the greater the recipient’s reliance upon income from the prospective capitation payment than upon income determined by the number of consultations delivered. The recipient now faces incentives to reduce costs in order to increase profitability — for example, producing fewer consultations or finding cheaper ways of producing them. Behavioural change ensues because undesirable behaviour is no longer rewarded (Milgrom and Roberts, 1992). Health care capitation contracts, as a form of
“supply-side cost sharing” (Ellis and McGuire, 1986), explicitly share financial risk otherwise borne by funders (for example, governments) and insurers with providers.

However, capitation contracts have complex effects on practitioner behaviour, as they share two specific types of financial risk between the purchaser and the provider: ‘controllable’ risk and ‘random’ risk. ‘Controllable’ risk relates to uncertainties of which the expected consequences can be anticipated in the aggregate, although are not predictable in respect of any specific occurrence. Cost consequences can be managed through the use of specific contract terms or institutional forms. Controllable risk is most efficiently managed when those capable of controlling its extent bear the financial costs of undertaking the undesirable actions or accrue the benefits of undertaking desirable actions. The specific controllable risk addressed by primary health care capitation is the provider propensity towards inefficient supplier-induced demand under subsidised fee-for-service remuneration (Zeckhauser, 1970). As providers bear at least some of the cost of their demand-raising choices, under capitation the number of unnecessary consultations reduces. Furthermore, providers are rewarded for engaging in consultation-reducing (and hence cost-reducing) preventative activities (Crampton, Sutton and Foley, 2001).

By contrast, ‘random’ risk relates to those factors for which the practitioner assumes either partial or complete financial responsibility via the risk-sharing contract, but is powerless to control. One such risk is an exogenous event causing an unpredictable and uncontrollable increase in demand for consultations (for example, a localised epidemic) — ‘exogenous’ risk. A second risk relates to differences between ex-ante anticipated demand for care within a population, and actual demand recorded ex post — ‘random demand variation’ risk. The aggregation of individuals’ demand risks into large insurance pools in order to reduce the cost of individual uncertainty by compensating from the pool in the event of the insured event materialising is an example of ‘controllable’ risk, more efficiently managed by an insurer with a diverse portfolio across which to spread the costs than by the risk-averse individual (Arrow, 1963). However, a discrepancy remains between anticipated demand, upon which premia or capitation payments are based, and actual demand which imposes actual costs.

When a large pool (for example, the population) is disaggregated into smaller pools (for example, patient lists), it is most unlikely that demand in each of the small pools will be the population average. At best, without explicit cream-skimming, half the pools will incur more demand than the population average, and half less. The smaller the size of the pool relative to the population, the greater will be the average absolute variation between the pool average upon which costs depend and the population average upon which remuneration is based — that is, the greater the extent of ‘random demand variation’ risk.
Capitation contracts let by large risk-pool managers (insurers and government funders) to a large number of smaller pool managers (service providers) fragment the efficiency-raising aggregate risk pool. The capitation contract transfers not just the amount of ‘controllable’ risk desired to be shared with the practitioner in order to alter the practitioner’s behaviour, but also a share of responsibility for ‘random demand variation’ and ‘exogenous’ risk. The United States Health Care Financing Administration considers capitated primary health care physician groups to be at substantial financial risk from random effects if they have fewer than 25,000 registered patients (Hagen, 1999).

On the one hand, capitation increases efficiency by mitigating the effects of unnecessary supplier-induced costs. On the other hand, by transferring responsibility for managing uncontrollable risk to more risk-averse practitioners, with less scope for diversification and efficient risk management than the funder/insurer (Milgrom and Roberts, 1992), efficiency is reduced. Providers’ incomes now become subject to factors over which they have no control or capacity to anticipate. In effect, providers assume an insurance role which they do not carry under fee-for-service contracts. The overall efficiency of capitation relative to fee-for-service depends upon whether the gains from improved management of ‘controllable’ risk exceed the losses from less efficient ‘random’ risk-management practices.

The effects of ‘random risk sharing’ on provider incomes may be either positive (that is, fewer consultations than anticipated/remunerated are provided) or negative (more consultations than anticipated/remunerated demanded) but, importantly, is beyond the provider’s control. The amount of risk shared is crucial for contract efficiency. The stronger the capitation contract incentive (that is, the higher is $f$ and the lower is $v$), the greater the proportion of ‘random’ risk that is shared, and the more the practitioner’s income comes to depend upon uncontrollable factors. If the ‘random’ effects are large compared to the ability to manage income via the ‘controllable’ factors, then irrespective of the amount of effort exerted pursuing the desired activities, the practitioner’s income becomes essentially a lottery. The incentive to pursue desirable behaviour is ‘crowded out’ by the random effects. Desired activities are not pursued. Rather, the practitioner will exert effort instead in activities that maintain or increase income given the amount of ‘random’ risk assumed (for example, ‘cream-skimming’ using information unknown to the funder to ensure that the patients for whom care management responsibility is assumed are more financially ‘desirable’) (Holmstrom and Milgrom, 1991).

United States evidence suggests that substantial changes in practitioner behaviour have been induced with only very weak capitation incentive contracts (Ma and Riordan, 2002), whereas only about 20 per cent to 25 per cent of the variation in individual demand for health care services can be predicted using
individual characteristics such as age, gender, income, ethnicity and past consumption of care (Robinson, 2004; Newhouse, 1996). These data suggest that there are very real ‘random’ financial risks for practitioners associated with using strong capitation incentives to manage ‘controllable’ risks when risk-adjusted fixed capitation payments are based on only a small range of individual demographic characteristics.

If the uncontrollable and unpredictable events were truly random, then over time, losses incurred by a provider in ‘bad’ years will be cancelled out by ‘profits in ‘good’ years. This is true of ‘exogenous’ risks. However, Howell (2007) posits that primary health care capitation contracts are exposed to serial correlation of profitability between years as a consequence of repeated transacting between a single provider and the same patients with persisting, but unknown (and hence uncompensatable via the prospective payment) risk factors that have ongoing effects upon the demand for care — that is, ‘random demand variation’ risk. The consequence is the emergence of habitually profitable and habitually loss-making practices. The correlation problem is further exacerbated by primary care practices generally being composed of only a small pool of individuals, where the persistent, atypical demand patterns of a few individuals (for example, underpinned by unknown and unknowable genetic predisposition) can have a significant effect on long-term practice profitability.

Correlation factors are less problematic in secondary and tertiary care provision as practitioner interaction with specific individuals tends to be episodic rather than ongoing, and the catchment from which demand derives is generally very much larger than a typical primary practice pool (Scott, 2000; McGuire, 2000). These effects are mitigated only by the recreation of larger pools amongst which to share the risks. If an individual general practitioner serves a group of typically between 1200 and 2000 patients, then based upon the United States evidence above, even with low-strength capitation incentives, financially viable primary care patient pools require the aggregation of the lists of between 12 and 21 practitioners. The greater the capitation incentive strength, the greater the number of practitioner lists required to be merged to counter the effect of ‘random’ risk sharing.

Figure 1 illustrates. Assume it costs an average practice $c$ to deliver an average primary care consultation (including all overheads and a fair return on the human capital and time invested by the practitioner) and that the practice delivers $q$ consultations. Average revenue received per consultation is $f/q + (v-c)$. Under a fee-for-service contract charged at cost ($f=0; v=c$), the practitioner makes no profits and no losses on all consultations delivered (that is, ‘breaks even’ at all values of $q$). The number of consultations delivered, $q$, is determined solely by consultation demand and the practitioner’s willingness to work. Under a capitation contract, however, the practitioner receives $v<c$ for each consultation.
delivered. The maximum number of consultations that the practitioner will deliver is $q=Q$, where the practice ‘breaks even’ financially. If demand arising from the population for which capitation is received results in fewer than $Q$ consultations being delivered, the practice makes a ‘windfall’ profit. If $q>Q$, the practice makes a financial loss.

A capitation contract incentivising desirable behaviours looks for providers delivering more than $Q$ consultations through their own cost-causing choices (that is, ‘controllable risk’) to reduce the number to $Q$ in order to remain financially viable, or for those with costs higher than $c$ to reduce them to $c$. However, the same contract will financially penalise those practitioners with costs $c$ and not over-producing, where random factors, rather than behavioural choices, lead to more than $Q$ consultations being delivered. If these practices are to remain financially viable, they must reduce costs below $c$ (for example, through shorter consultations), ration services (for example, institute waiting lists) or pass the extra costs on in some other way (for example, institute a patient payment $y$ in addition to the capitation contract payment $v$, which is typically paid by the insurer or funding body).

**Figure 1: Average Revenue Per Consultation: Standard Capitation**

The sharper the capitation contract incentive (that is, the lower is $v$ and the higher is $f$), the steeper is the slope of the average revenue curve (Figure 1), the greater the profits and losses, and the greater the additional costs that must be borne by the patients of ‘unprofitable’ practices ($q>Q$). Ironically, ‘unprofitable’ practices meeting all ‘controllable’ risk expectations make losses in the first place because their patient base has higher demand (that is, is ‘sicker than average’). Capitation contracts result in sicker-than-average patients bearing more of the consequences of ‘random’ risk sharing, in either lower care quality (shorter consultations, waiting lists) or higher prices ($y$ charged to them) than the ‘healthier-than-average’ patients of ‘profitable’ practices.
Figure 2 illustrates the effect of ‘unprofitable’ practices being able to charge patients $y$ to recover ‘random’ risk costs. Assume that the ‘most unprofitable’ practitioner not over-producing at cost $c$ must deliver $Q_1$ consultations to meet demand at $f$ and $v$. By charging patients $y$ per consultation to break even, the practice will still produce $Q_1$ consultations (technically, the practice’s average revenue curve moves upward). As primary care practitioners have some market power due to product differentiation arising from patient preferences for the attributes of individual practitioners and repeated transacting between the same individuals (Scott, 2000; Dranove and Satterthwaite, 2000), within bounds such price increases can be undertaken without invoking the loss of large numbers of patients to other practices. However, if one practice can charge a patient fee $y$ without losing patients (or the failure to charge $y$ has no substantial effect upon a given practitioner’s demand), all other ‘profitable’ and ‘less unprofitable’ practices will also be able to charge $y$. ‘Unprofitable’ practices delivering between $Q$ and $Q_1$ consultations now make profits instead of losses, and the ‘profitable’ practices producing fewer than $Q$ consultations make higher profits than before. The total number of consultations produced is higher than anticipated by the capitation contract based upon remuneration from $f$ and $v$ alone.

**Figure 2: Average Revenue Per Consultation: Patient pays $y$**

If practitioners can charge patients $y$, there is no need to engage in cost- and service quality-reducing activities such as rationing or shortening consultations (these actions are commonplace where patient charging is prohibited; for example,
in England’s NHS). However, all patients face an additional financial cost in lieu of the quality reduction that would otherwise be borne by patients of ‘unprofitable’ practices, regardless of whether their practice would have been required to engage in such activities in order to break even. All practices now make profits, with those facing least demand becoming substantially more profitable than if patient charging was prohibited.

Furthermore, the ability to charge patients eliminates the justification for using capitation to alter practitioner behaviour in the first place. If practices can levy charges to cover costs of ‘random’ risk allocation, they can also levy charges to cover the additional costs arising from their ‘controllable’ risk choices. There is no financial penalty from engaging in the delivery of over-many consultations. The number of consultations delivered returns to the level under fee-for-service remuneration, but the total cost of delivering those consultations increases as a consequence of practitioners reaping profits from random demand variation and patient charging that they were unable to appropriate under fee-for-service remuneration. Despite raising the costs of service delivery, the contracts are impotent in constraining practitioner supply. Undesirable practitioner behaviour must now be controlled using costly, overt means such as direct observation or regulation. Ironically, capitation is widely used precisely because it is more cost-effective than monitoring where the desired behaviour is either extremely costly or infeasible to directly observe (such as in third-party purchasing — Milgrom and Roberts, 1992).

Systems enabling patient charging (\(y\)) thus result in more costly consultations. They also invoke perverse distributional consequences. The higher costs of risk management are borne only by individuals consuming consultations — that is, the sick. The ‘sicker’ the patient (that is, the more consultations consumed) the greater the contribution towards the higher risk management costs. ‘Healthy’ patients (that is, those consuming no consultations) pay none of the inflated risk-management costs imposed by system design. Patient payment systems thus allocate the higher-than-expected costs in the form of a perfectly risk-rated ‘premium’ per consultation based upon patient health state (or equivalently, as a ‘consumption tax’ imposed on the sick). The ‘well’ are rewarded for their ‘good’ health state by not being required to pay any of the risk-management costs shared with practitioners via capitation contracts and subsequently ‘passed on’ to the sick. Such arrangements are particularly antithetic to the principles of socially motivated insurance schemes, where patient income (via taxation), rather than health state (via patient payments per consultation required), is the preferred metric via which the financial costs of the scheme are allocated.

**The New Zealand Primary Health Care Strategy**

The NZPHCS was introduced in July 2002 (King, 2001). Capitation payments are made to Primary Health Organisations (PHOs), which then contract with primary
care providers to deliver services to patients. Practically all primary care providers are paid via a ‘back-to-back’ contract whereby the capitated ‘GMS/Nurse’ payment is ‘passed through’ by PHOs directly to providers, who are predominantly sole practitioners with lists of between 1200 and 2000 patients (Howell, 2005). Primary objectives of the strategy are to increase the share of government funding from between 30 per cent (Austin, 2004) and 40 per cent (King, 2001) in 2001 to around 80 per cent (Howell, 2007a), and the promotion of preventative care. The capitation formulae are broadly risk-rated, using patient age, gender and past consumption of care characteristics, and average practice ethnicity and income (determined by the deprivation quintile of the geographical area in which the practice is located) characteristics.

Service Providers Set and Charge v plus y

The first defining feature of the NZPHCS is that, although the government share of primary-care sector costs is increasing, patients are still required to pay direct to service providers the balance of costs of primary care delivery not covered by the government payments. That is, whilst the government sets and pays \( f \), each service provider sets \( v \) individually, and the patients pay it. This charging arrangement arises from ‘grandfathering’ through into the NZPHCS the terms of the 1938 agreement between the government and medical practitioners enabling practitioners to charge patients directly for any costs not covered by government subsidies (Ashton, 2005; Howell, 2005). The ‘grandfathering’ makes no distinction between the risk-sharing imposed under capitation and the ‘risk-free’ status of practitioners under fee-for-service.

Section 1 suggests that provider setting of \( v \) (and \( y \) if losses arise) eliminates any possibility of using capitation to alter provider behaviour. Whilst patient out-of-pocket costs may decrease due to increased government payments, total costs for the same number of consultations delivered will increase as costs of risk management increase. The additional costs will be disproportionately borne by those consuming higher numbers of consultations.

In the three years following the policy’s introduction, government primary care funding increased by 43 per cent (Hefford, Crampton and Foley, 2005). However, the government’s total share of expenditure did not increase by as much as the Minister had expected given the substantial additional injection of government funding (King, 2004). Figure 2 suggests the most likely explanation is that the strategy confers the ability for loss-making providers to ‘pass on’ to patients the costs of both ‘controllable’ and ‘random demand variation’ risk. Evidence that loss-making practices are likely avoiding financial distress by passing on risk-management costs is provided by McAvoy and Coster (2005:11), who report that ‘many GPs have already benefited, reporting improved financial incomes and financial benefits from joining PHOs’. This observation supports the contention that practices previously only ‘breaking even’ under
fee-for-service remuneration are now making windfall profits, and that some may be using the latitude provided by unprofitable providers increasing patient charges to extend the scope of their profitability.

Further indication that the cost of consultations is likely higher under the NZPHCS than under fee-for-service is contained in CBG’s (2004) observation that providers not yet receiving capitation payments for patients were charging less on average for consultations for identical patient groups than those providers receiving capitation subsidies. CBG (2004) finds this observation puzzling, as capitated providers receive, on average, a substantially higher ‘notional averaged treatment subsidy’ per consultation than those providers remaining on fee-for-service payment. Whilst it might be possible to explain the observation by patient self-selection to capitated cover on the basis of higher individual health need, under the NZPHCS it is practitioners, not patients, who select the remuneration form of the benefits paid in respect of their registered patients. Practitioner self-selection to a payment system that, despite receiving higher government payments per patient than previously, still sees them levying patient charges that are higher than those of less-well subsidised fee-for-service practitioners, is quite consistent with Figure 2’s contention that the cost of service delivery is higher under capitation.

Indeed, if those practices likely to become unprofitable under the NZPHCS due to the high demands of their patient base were the ones which rationally refrained from opting into the capitated system, and those with low demands self-selected to capitation because they are able to access profits not available otherwise, the observation would appear to confirm that both ‘profitable’ and ‘unprofitable’ capitated practices are charging additional fees to cover losses imposed upon ‘unprofitable’ practices by the changes in risk allocation. If, perchance, it was high-demand practices that opted disproportionately for capitation (for example, improved cash flows under capitation would reduce costs associated with the collection of bad debts), CBG’s (2004) observation is still consistent with Figure 2. It simply confirms that unprofitable practices delivering more than Q consultations are forced to increase fees under capitation in order to break even, thereby raising the average fees charged by capitated practices above the level charged by uncapitated practices that face no losses at all for consultations provided beyond Q as they are remunerated for each consultation at its cost c.

Increasing Incentive Strength Over Time

The second defining feature of the NZPHCS is that government capitation payments have increased over time in accordance with a set of political priorities determined by government budgets, rather than any systematic consideration of the ability of the capitated entities to manage the amount of risk shared. The first beneficiaries were individuals aged over 65 and under five, and those
patients registered in PHOs where more than 50 per cent of the registered individuals were in the lowest income quintiles or of Maori or Polynesian ethnicity (“Access Practices” — the remainder are termed “Interim Practices”). Over subsequent years, higher subsidies were extended to five–14-year-olds, 15–24-year-olds and 45–64-year-olds. The final group, 26–44-year-olds, received higher subsidies from 1 July 2007. Whilst many of the differences between Access and Interim practice payments have now been reduced, small differences remain for children under 14 across all practices, and substantial differences remain in respect of 25–44-year-old patients of practices not yet deemed to have qualified to receive increased funding.²

Each increase in government funding has been associated with political expectations that increases in the ‘notional averaged treatment subsidy’ embodied in the capitation payment will be directly matched by corresponding reductions in the fees charged by practitioners for subsidy classes receiving more generous government assistance. For example, the Prime Minister stated that the 1 July 2006 capitation increases meant that “700,000 people aged between 45 and 64 would now pay $27 less for doctor visits and $3 for prescriptions — instead of $15 — if they enrolled with a primary health organisation”.³ Capitation payments are routinely referred to as “subsidies to lower the cost of doctors’ visits”.⁴ This obfuscates the significant changes that have occurred in risk allocation. Despite substantial risk-bearing distinctions, prospective capitation payments appear to be treated as directly equivalent to the risk-free (for the practitioner) ex-post treatment subsidies paid under the preceding system.

Clear expectations exist that some patients will be charged different prices as a consequence of their subsidy class, even when all population groups receive capitation subsidies. For example, there is strong pressure for practitioners not to charge a fee for consultations delivered to children under five years old. It is also expected that individuals with high-use status will pay lower out-of-pocket charges than an otherwise identical patient without high-use status, even though the high-use premium is paid as a capitated sum rather than a per-visit subsidy. DHB and MoH policy material supports the contention that differential approaches to fee-setting based upon the extent of the subsidies received for patients of different classes will prevail in the event of price monitoring and, if necessary, price regulation occurring (DHB, 2006).

The overall effect of the staged rollout is that, over time, the average share of a practitioner’s income that comes from government subsidies (f) increases and the share from patient payments (v) decreases. The capitation incentive strength thus becomes increasingly sharp, as per Figure 1. At each increase, the

² http://www.moh.govt.nz/moh.nsf/indexmh/phcs-funding
³ http://www.stuff.co.nz/stuff/0,2106,3719646a7144,00.html
⁴ http://www.moh.govt.nz/moh.nsf/indexmh/phcs-funding
share of both ‘predictable’ and ‘random’ financial risk that practitioners are required to bear becomes greater. There is no actuarial or cost-benefit evidence provided that it is reasonable to expect providers to efficiently manage either ‘controllable’ or ‘random’ risks of the magnitude shared by the contracts (Howell, 2007a). Monopsony power held by the government purchaser gives service providers few options but to accept the capitation contract terms, as the difference in subsidy extent between the NZPHCS and the alternative regimes is so large that patients will almost certainly defect to other providers if their practitioner does not ‘sign up’ (Howell, 2005). It is therefore highly likely that the NZPHCS arrangements will lead to unanticipated outcomes.

Given the very small size of the patient pools and increasing incentive contract strength, even higher patient charges and increases in practice profitability will occur when subsidy rollout is complete than observed by CBG (2004) and McAvoy and Coster (2005) initially. It has become increasingly more costly with each addition to the subsidised population to deliver the same number of consultations as provided under the fee-for-service regime. The ability to levy charges to maintain practice profitability disincentivises practitioners from undertaking otherwise loss-minimising, efficiency-increasing mergers of their patient lists to manage the increased share of random risk they must now bear. The normal institutional responses that can be expected in capitated systems where the funder sets and pays both f and v are negated by NZPHCS institutional design. Despite the creation of PHOs to aggregate practitioner activity, and the potential to manage random risk more efficiently by list aggregation, the ownership of the patient list, and receipt of capitated patient payments, continues to remain with individual practitioners (Howell, 2005).

**Distributional Implications**

The inevitable consequence of the NZPHCS design is that the higher-than-anticipated costs are being borne only by those consuming consultations — that is, the sick — in proportion to the number of consultations consumed. Those consuming no consultations bear none of the additional costs, but instead benefit from the expectation that a consultation, if required, will cost less out-of-pocket than previously.

The inequity occurs because, whereas in a typical managed care or social insurance system the full anticipated costs of treatment would be collected ex ante from all insured individuals, in New Zealand, only the 80 per cent funded by the government is collected in this manner via taxation. As the government’s payments are fixed, it bears none of the costs of random demand variation. The entity best placed, given its ability to spread costs across a large population, to underwrite individual demand variation risk, has abrogated all risk-management responsibility to very much smaller practitioner entities much less able to spread risk-management costs efficiently. Patient payments now potentially and actually
vary substantially between practices principally because of the varying demands of a small number of individuals registered at the same practice, rather than the non risk-related costs of delivering that care.

That the demands of other patients registered at the same practice has now become a significant determinant of individual cost appears quite antithetic to policy aspirations that costs would be shared in a manner independent of patient health state. Whilst individual health state determines the number of contributions made by an individual towards the additional risk costs, the size of those payments is likely determined by the profitability of the practice at which the patient is registered. Inequity arises even if profitable practices do not take advantage of the ability to raise prices alongside unprofitable ones. If the practice is profitable because average demand is low, and the practitioner is altruistic and opts not to charge a payment \( y \), or even charges a payment \( v \) that is lower than that upon which capitation averages is based, then it is fewer sick patients who will pay lower fees. However, if the practice is unprofitable due to a sicker-than-average patient base, the sicker-than-average patients will face not only the standard capitation payment \( v \), but \( y \) as well. System design makes it inevitable that resources will be allocated inequitably, to the disadvantage of those who are sicker than average.

**Illustration by Stylised Examples**

This section uses stylised examples based on actual events to illustrate the effects of the allocation of random risk under the NZPHCS. The first example is a strike by hospital junior doctors; the second is a decision by District Health Boards to cull hospital waiting lists by removing all patients who would not be seen by hospital staff within six months. In both cases, unmet hospital demand was referred back to primary care practitioners for management, causing primary care demand to increase, as each patient referred back would require at least one, and possibly more, additional consultations not anticipated in the capitation formula design. Whilst the shocks are exogenous, and ultimately the additional demand could be factored into revised capitation payments, the examples also highlight how different allocation of individuals of different health states amongst practices in the presence of variable practitioner-set patient payments leads to higher payments for the patients of practices with sicker patients facing higher demands.

The examples highlight the fallacy of equating the ‘notional averaged treatment subsidy’ under capitation to the risk-free treatment subsidy under the previous fee-for-service system. Under fee-for-service, practitioners would be indifferent to any demand shock, as each extra consultation delivered would be remunerated at its full cost \( c \). Under capitation, however, a practice incurs the full costs of an additional consultation, but receives no additional government payments. As the capitated patient is charged a fee less than cost \( (v < c) \), the
additional income from the additional consultations will be insufficient to cover their additional, and unanticipated cost. The practice now incurs a deficit in respect of the additional demand. As practitioners can charge patients, it is presumed that the deficit is passed on in higher patient fees \((\text{patient payment} = v + y)\).

**The Base Case**

A simple numerical example illustrates. Suppose each consultation costs the practice $50, and the patient payment \((v)\) is initially set at $10. Each additional consultation arising from a DHB referral imposes a $40 deficit on the practice. The more patients referred back by the DHB and the greater the number of consultations required for each patient, the greater the deficit incurred. If the practice is unable to charge the referred patients the full cost of the additional consultations, to break even the additional costs must be levied to all patients via an increased patient payment \(y\). All patients, including those not receiving the additional consultations, now pay the risk premium created by an unpredicted and unpredictable (that is, ‘random’) change in the demand of a small number of patients registered at the practice. As the patient list is the risk pool, prices faced by those patients in the pool whose demand does not change are determined by those whose demand has changed.

Assume now that two practices have otherwise identical costs and patient lists with identical ex-ante characteristics, but one has 10 patients referred back and the other has 20 (that is, its patients are ‘sicker’ but this is not recognised in the capitation formula). The deficit incurred by the second practice is twice that of the first practice. The patients of the second practice, who are sicker on average, will face payment increases twice the size of the first. Patients with identical capitation status now pay different prices depending upon the practice at which they are registered, even though both practices have the same service delivery costs.

**Different Capitation Classes and Patient Charges**

Now consider the effects of the mandatory requirement that individuals for whom higher capitation payments are received when well must also pay commensurately lower patient payments when they seek a primary care consultation. Assume the base-case practice charges nothing to high-capitated patients under five years old and $30 per consultation to low-capitated 25–44-year-old patients. The deficit for each additional DHB-referred under-five consultation is $50 and for each additional 25–44 consultation $20. The higher the proportion of higher-capitated (ex-ante assessed as ‘high needs’ or ‘high political priority’) patients referred back by the DHB, the greater the deficit incurred and the higher the commensurate price increases to all patients of the practice must be. The patients of practices with large numbers of higher-capitated
patients referred back will pay proportionately more, because the practice is obligated to charge the higher-capitated patients lower fees.

The costs of the DHB waiting list cull are unlikely to be trivial. Typically, higher-capitated (and politically favoured) elderly and young people are disproportionately represented in hospital waiting lists. Thus, the probability of a waiting-list patient referred being a high-capitation individual will be substantially higher than the probability of the patient being a low-capitation individual. These individuals are also likely to require multiple additional primary-care consultations in order to continually treat the waiting-list condition and to reassess the eligibility for re-entry into the waiting-list system. These non-trivial additional costs, imposed by DHBs shifting demand and financial risk into the primary sector, will have a non-trivial effect upon the costs paid by all sick individuals seeking primary care.

This example illustrates the fallacy of treating the capitation subsidy as if it is a ‘notional averaged treatment subsidy’ when setting \( v \) for each subsidy group. Patients from two different subsidy categories may each have identical needs for treatment, but requiring each to pay different amounts as each receives a different capitation subsidy confuses the payment of an ex-ante risk-rated insurance premium subsidy with a politically motivated wealth transfer. The former approach compensates the risk manager for the extra costs anticipated in respect of individuals with higher demands, and typically means that no distinction needs to be made between patients of any premium class when treatment is sought ex post. The latter approach echoes the pre-NZPHCS arrangements, when different subsidies were paid for different classes of individual for politically motivated wealth-transfer reasons.

By confusing differences in anticipated demand for care with politically motivated wealth transfers under the NZPHCS, sick patients of practices with greater exposure to unanticipated demand shocks associated with individuals of the more favoured group end up contributing not just to the costs of demand variation in their own payments for health care, but also a portion of politically motivated wealth transfer because the two functions are bundled together in the one patient payment. Such transfers are avoided in typical social insurance systems when the wealth transfer is achieved using a premium subsidy. Less politically favoured groups pay more of the costs of the system by paying a higher premium top-up ex ante (for example, in New Zealand’s workers Accident Compensation (ACC) system, employers pay a premium for employees based upon industry risk characteristics, and the employee’s contribution is paid as a proportion of taxable income). This eliminates any need to adjust payments to achieve wealth transfers when treatment is sought (under the ACC system, there is no difference in patient payments based upon income or subsidy class).
Individuals with equal need of treatment pay equal amounts at the point of service delivery.

To avoid distortions of the kind illustrated in the example, either premium top-ups should be charged ex ante in respect of each individual in accordance with political wealth-transfer motivations and identical payments levied ex post; or identical payments should be levied to all patients ex post and an additional fee-for-service government subsidy paid in respect of each treatment delivered to the member of a politically favoured group, as occurred pre-NZPHCS. Under such arrangements, there is separation and transparency between the consequences of risk-management practices and wealth transfers, and the sick are not required to pay an additional consumption tax on primary health care in order to further political wealth-distribution goals.

**Patient List Mix**

Assume now that one practice has a high proportion of higher-capitated individuals ‘on the books’ (Practice A), and another practice (Practice I) has a low proportion. Practice A thus receives a higher proportion of its revenue from fixed payments, and it can only charge its patients on average low values of $v$. Practice I receives a lower proportion of its income from fixed payments, and more from patient payments (that is, $v$ is higher). The capitation contract incentive is thus higher for Practice A.

Both practices provide the same number of consultations ($K$) in a given period. The average patient payment $v$ at Practice A is $20 per consultation, and at Practice I $35. Assume also that, as a consequence of the DHB patient referral, each is required to provide an additional 200 consultations in a given period. Practice A incurs a deficit of $200 \times 30 = 6000$, and Practice I a deficit of $200 \times 15 = 3000$. If each practice spreads the additional costs across all patients, Practice A patients will pay an additional $6000/(K+200)$, twice the increase faced by Practice I patients, $3000/(K+200)$. Thus, DHB referral results in the ex-ante assessed ‘higher need’ Practice A patients, who are deemed less able to meet the costs of higher patient payments, and more likely to be dissuaded from seeking primary care by the size of the payment, and ostensibly more likely to need care in the first place, facing higher patient payment increases than the ‘lower priority’ Practice I patients.

This example illustrates numerically the increasing costs of uncontrollable events under sharper incentives. It also illustrates the inequitable consequences of confusing a premium subsidy with a treatment subsidy in the absence of consideration of the locus of residual financial risk-bearing.
Regulatory Intervention in Practice Price Setting

Inevitably, patient payments have become more variable between practices as the proportion of the population eligible for capitation subsidy has increased. Consequently, the likelihood of DHBs invoking their NZPHCS price-regulation powers has increased. However, if prices are regulated as if the patient payment is a risk-free ‘notional averaged treatment subsidy’, practice financial viability may be severely compromised. DHB materials and politicians’ expressed intentions suggest that neither group fully appreciates the extent to which the NCPHCS has shifted substantial amounts of additional random-risk costs onto practices relative to the pre-2002 system. Consequently, fears expressed by practitioners about their fees becoming subject to price regulation by DHBs under the currently voiced understandings are substantiated.

Assume that a naïve regulator, knowing that Practice A receives higher capitation payments, and therefore, in line with treatment benefit pass-through expectations, must charge lower patient out-of-pocket payments than Practice I, is faced with the price rises in the example above. If the regulator presumes that the capitation payment is simply a ‘notional averaged treatment subsidy’, then a price increase of $6000/($K + 200) imposed by Practice A is ‘unreasonable’ given that Practice I with the same demand increase imposes a price increase of only $3000/($K + 200). If the regulator restricts the price increase by Practice A to that imposed by Practice I, then Practice A becomes financially unviable.

Effective regulation of capitated practices requires specialist insurance and risk-management knowledge. Such regulation is problematic even in countries where there is considerable experience in this form of regulation (Hagen, 1999). Given the added complications from the bundling of wealth transfers with patient payments, effective regulation within the NZPHCS would appear to be an extremely complicated, costly and risky endeavour.

Conclusion

Using theory and stylised examples, this paper illustrates the perverse consequences that can arise when the design of capitation contracts does not give due consideration to the consequences of sharing both ‘controllable’ and ‘random’ risk with service providers. The NZPHCS capitation contracts, sharing very large amounts of random risk with very small risk pools, and allowing patient charges to be set by practitioners, negate the rationale for using capitation to induce behavioural changes in providers, increase the cost of providing consultations, and allocate the additional costs disproportionately to those who consume services most. Whilst increased government funding has reduced out-of-pocket costs for patients, relative to its fee-for-service predecessor the

---

NZPHCS is operating substantially less efficiently. The explanation is likely inadequate consideration in its design given to the optimal allocation of financial risk. Unless consideration is given to the extent that providers are required to act as insurers of patients on their lists, inequitable and over-costly outcomes will persist.

References


